

Linear Regression:
Mis-specification and Weighted
Least Squares

The Linear Model: recap

For a single observation

- Y scalar
- $\mathbf{x}$ is $1 \times p$
- β is $p \times 1$

We assume that

$$\mathbb{E}[Y|\mathbf{x}] = \mathbf{x}\beta = \sum_{j=1}^p x_j \beta_j$$

and

$$\mathbb{V}\text{ar}[Y|\mathbf{x}] = \sigma^2$$

The Linear Model: recap (cont.)

For a data set of size n comprising

- outcome data $\mathbf{y} = (y_1, \dots, y_n)^\top$
- predictor data $\mathbf{X}$ – an $n \times p$ matrix

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{bmatrix}$$

we assume

$$\mathbb{E}[\mathbf{Y}|\mathbf{X}] = \mathbf{X}\beta \quad \mathbb{V}\text{ar}[\mathbf{Y}|\mathbf{X}] = \sigma^2 \mathbf{I}_n$$

The Linear Model: recap (cont.)

This is equivalent to the model

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

where ε is an $(n \times 1)$ vector of random variables with

$$\mathbb{E}[\varepsilon|\mathbf{X}] = \mathbf{0}_n \quad \mathbb{V}\text{ar}[\varepsilon|\mathbf{X}] \equiv \mathbb{E}[\varepsilon\varepsilon^\top|\mathbf{X}] = \sigma^2\mathbf{I}_n$$

The Linear Model: recap (cont.)

With $\mathbf{X}$ fixed, estimate parameters β and σ^2 using *ordinary least squares* (OLS)

$$\hat{\beta} = \arg \min_{\beta} (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta)$$

that is, $\hat{\beta}$ solves

$$\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\beta) = \mathbf{0}_p$$

so that

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

Note that

$$\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\beta) = \sum_{i=1}^n \mathbf{x}_i^\top (y_i - \mathbf{x}_i\beta)$$

showing the form of the *estimating function*.

The Linear Model: recap (cont.)

By standard theory, we have

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

with corresponding *estimator*

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

The Linear Model: recap (cont.)

For estimating σ^2

$$\hat{\sigma}^2 = \frac{1}{n-p} (\mathbf{Y} - \mathbf{X}\hat{\beta})^\top (\mathbf{Y} - \mathbf{X}\hat{\beta})$$

is the usual estimator. We may write

$$\mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = \mathbf{H}\mathbf{Y}$$

say, where

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$$

is a symmetric $n \times n$ matrix.

The Linear Model: recap (cont.)

Thus

$$\begin{aligned}(\mathbf{Y} - \mathbf{X}\hat{\beta})^\top (\mathbf{Y} - \mathbf{X}\hat{\beta}) &= (\mathbf{Y} - \mathbf{HY})^\top (\mathbf{Y} - \mathbf{HY}) \\ &= \mathbf{Y}^\top (\mathbf{I}_n - \mathbf{H})\mathbf{Y}.\end{aligned}$$

Note that by symmetry

$$\mathbf{H}^\top \mathbf{H} = \mathbf{H}\mathbf{H} = \mathbf{H}\mathbf{H}^\top$$

and also

$$\mathbf{H}^\top \mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top = \mathbf{H}.$$

The Linear Model: recap (cont.)

Unbiasedness: Taking expectations with respect to the presumed data generating model for $\mathbf{Y}$, where the true value of the parameter is β , we have that

$$\begin{aligned}\mathbb{E}[\hat{\beta}|\mathbf{X}] &= \mathbb{E}\left[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \middle| \mathbf{X}\right] \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}[\mathbf{Y}|\mathbf{X}] && \text{as } \mathbf{X} \text{ is constant} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\beta) \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{X})\beta \\ &= \beta.\end{aligned}$$

The Linear Model: recap (cont.)

Variance:

$$\begin{aligned}\mathbb{V}\text{ar}[\widehat{\beta}|\mathbf{X}] &= \mathbb{V}\text{ar}\left[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \middle| \mathbf{X}\right] \\&= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{V}\text{ar}[\mathbf{Y}|\mathbf{X}] \left\{ (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \right\}^\top \\&= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \sigma^2 \mathbf{I}_n \left\{ \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \right\} \\&= \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{X}) (\mathbf{X}^\top \mathbf{X})^{-1} \\&= \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}.\end{aligned}$$

The Linear Model: recap (cont.)

- line 2 follows by the properties of variance-covariance matrices for multivariate random variables: for constant matrix $\mathbf{A}$,

$$\text{Var}[\mathbf{A}\mathbf{Y}] = \mathbf{A}\text{Var}[\mathbf{Y}]\mathbf{A}^\top.$$

- line 3 follows as

$$\mathbf{X}^\top \mathbf{X}$$

is symmetric.

Suppose we have two sets of predictors that yield matrices

$$\mathbf{X}^{(1)} : n \times (p - r)$$

$$\mathbf{X}^{(2)} : n \times r$$

Let $\mathbf{X} = [\mathbf{X}^{(1)} \ \mathbf{X}^{(2)}]$ be the combined $n \times p$ matrix, and suppose

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \equiv \mathbf{X}^{(1)}\beta^{(1)} + \mathbf{X}^{(2)}\beta^{(2)} + \epsilon$$

with β , $\beta^{(1)}$ and $\beta^{(2)}$ are p , $p - r$ and r -dimensional parameter vector and sub-vectors respectively.

- **True Model:**

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon = \mathbf{X}^{(1)}\beta_{TRUE}^{(1)} + \mathbf{X}^{(2)}\beta_{TRUE}^{(2)} + \epsilon$$

- **Fitted Model:**

$$\mathbf{Y} = \mathbf{X}^{(1)}\beta^{(1)} + \epsilon.$$

For the estimators from the fitted model, we have

$$\hat{\beta}^{(1)} = (\{\mathbf{X}^{(1)}\}^\top \{\mathbf{X}^{(1)}\})^{-1} \{\mathbf{X}^{(1)}\}^\top \mathbf{Y} = \mathbf{A}^{(1)} \mathbf{Y}$$

say, and

$$\hat{\sigma}_{(1)}^2 = \frac{1}{n - (p - r)} \mathbf{Y}^\top (\mathbf{I}_n - \mathbf{H}^{(1)}) \mathbf{Y}$$

The estimator $\widehat{\beta}^{(1)}$ is in general *biased*:

$$\begin{aligned}\mathbb{E}[\widehat{\beta}^{(1)}|\mathbf{X}] &= \mathbf{A}^{(1)}\mathbb{E}[\mathbf{Y}|\mathbf{X}] = \mathbf{A}^{(1)}(\mathbf{X}^{(1)}\beta_{TRUE}^{(1)} + \mathbf{X}^{(2)}\beta_{TRUE}^{(2)}) \\ &= \mathbf{A}^{(1)}\mathbf{X}^{(1)}\beta_{TRUE}^{(1)} + \mathbf{A}^{(1)}\mathbf{X}^{(2)}\beta_{TRUE}^{(2)} \\ &= \beta_{TRUE}^{(1)} + \mathbf{A}^{(1)}\mathbf{X}^{(2)}\beta_{TRUE}^{(2)}\end{aligned}$$

as

$$\mathbf{A}^{(1)}\mathbf{X}^{(1)} = (\{\mathbf{X}^{(1)}\}^\top \{\mathbf{X}^{(1)}\})^{-1} \{\mathbf{X}^{(1)}\}^\top \mathbf{X}^{(1)} = \mathbf{I}_{p-r}.$$

Thus

$$\mathbb{E}[\widehat{\beta}^{(1)}|\mathbf{X}] \neq \beta_{TRUE}^{(1)}$$

in general.

We have equality if and only if

- *either* $\beta_{\text{TRUE}}^{(2)} = \mathbf{0}_r$ (i.e. the $\mathbf{X}^{(2)}$ predictors are not influential)
 - ▶ in this case, the model is correctly specified
- *or* the predictors in $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are *orthogonal*, as then

$$\mathbf{A}^{(1)}\mathbf{X}^{(2)} = \mathbf{0}_{p-r,r}$$

or equivalently

$$\{\mathbf{X}^{(1)}\}^\top \{\mathbf{X}^{(2)}\} = \mathbf{0}_{p-r,r}.$$

We also have by standard theory, for the fitted model,

$$\mathbb{V}\text{ar}[\hat{\beta}^{(1)}|\mathbf{X}] = \sigma^2(\{\mathbf{X}^{(1)}\}^\top \{\mathbf{X}^{(1)}\})^{-1} = \sigma^2 \mathbf{S}_{11}^{-1}$$

say, whereas if the correct full model is fitted, we have

$$\begin{aligned}\mathbb{V}\text{ar}[\hat{\beta}|\mathbf{X}] &= \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1} \\ &= \sigma^2 \begin{bmatrix} \{\mathbf{X}^{(1)}\}^\top \{\mathbf{X}^{(1)}\} & \{\mathbf{X}^{(1)}\}^\top \{\mathbf{X}^{(2)}\} \\ \{\mathbf{X}^{(2)}\}^\top \{\mathbf{X}^{(1)}\} & \{\mathbf{X}^{(2)}\}^\top \{\mathbf{X}^{(2)}\} \end{bmatrix}^{-1} \\ &= \sigma^2 \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix}^{-1}\end{aligned}$$

say.

Over-simplification (cont.)

On inverting the block matrix, we have for the variance covariance block for the $\beta^{(1)}$ component

$$\sigma^2(\mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21})^{-1}.$$

For square matrices $\mathbf{A}$ and $\mathbf{B}$, we write

$$\mathbf{A} \geq \mathbf{B}$$

to mean that $\mathbf{A} - \mathbf{B}$ is positive semi-definite. Thus, in general

$$(\mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}) \leq \mathbf{S}_{11}$$

as $\mathbf{S}_{22}$ is positive definite, so that

$$(\mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21})^{-1} \geq \mathbf{S}_{11}^{-1}.$$

Thus the variance of the estimator for $\beta^{(1)}$ from the full model is *larger* than that for the reduced model; we have equality if

$$\mathbf{S}_{12} = \{\mathbf{X}^{(1)}\}^\top \{\mathbf{X}^{(2)}\} = \mathbf{0}_{p-r,r}$$

that is, in $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are *orthogonal*.

However, recall that the estimator is *biased*; a fairer comparison involves using the *mean-squared error* (MSE) which is the sum

$$\text{Var}[\hat{\beta}^{(1)}|\mathbf{X}] + (\mathbb{E}[\hat{\beta}^{(1)}|\mathbf{X}] - \beta^{(1)})(\mathbb{E}[\hat{\beta}^{(1)}|\mathbf{X}] - \beta^{(1)})^\top$$

For the full model, the estimator is unbiased, so the MSE is merely

$$\sigma^2(\mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21})^{-1}$$

whereas for the fitted model, the MSE is

$$\sigma^2\mathbf{S}_{11}^{-1} + \mathbf{A}^{(1)}\mathbf{X}^{(2)}\beta^{(2)}\{\beta^{(2)}\}^\top\{\mathbf{X}^{(2)}\}^\top\{\mathbf{A}^{(1)}\}^\top$$

and so which of the two MSEs is larger depends on the magnitude of $\beta^{(2)}$.

The estimator $\hat{\sigma}_{(1)}^2$ has expectation

$$\sigma^2 + \frac{1}{n - (p - r)} \{\mathbf{X}^{(2)}\}^\top \{\beta^{(2)}\}^\top (\mathbf{I}_n - \mathbf{H}^{(1)}) \mathbf{X}^{(2)} \beta^{(2)}$$

so there is a *positive bias*.

- **True Model:**

$$\mathbf{Y} = \mathbf{X}^{(1)} \beta_{\text{TRUE}}^{(1)} + \epsilon;$$

- **Fitted Model:**

$$\mathbf{Y} = \mathbf{X}^{(1)} \beta^{(1)} + \mathbf{X}^{(2)} \beta^{(2)} + \epsilon.$$

where we know that the true value of $\beta^{(2)} = \mathbf{0}_r$.

Standard theory applies:

- the least squares estimator is *unbiased*
 - ▶ $\hat{\beta}^{(2)}$ is unbiased for the true value $\beta_{TRUE}^{(2)} = \mathbf{0}_r$.
- the estimator has variance

$$\sigma^2(\mathbf{X}^\top \mathbf{X})^{-1} = \sigma^2 \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix}^{-1}$$

The variance of $\hat{\beta}^{(1)}$ from the fitted model is

$$\sigma^2(\mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21})^{-1}$$

which, as before, is no smaller than $\sigma^2\mathbf{S}_{11}^{-1}$.

Thus the variance is larger than the model that does *not* fit spurious variables.

We generalize the model

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

so that

$$\mathbb{E}[\mathbf{Y}|\mathbf{X}] = \mathbf{X}\beta \quad \mathbb{V}\text{ar}[\mathbf{Y}|\mathbf{X}] = \mathbf{W}^{-1}$$

for a fixed, positive definite symmetric matrix $\mathbf{W}$.

Weighted least squares (cont.)

This is equivalent to a model where ε is an $(n \times 1)$ vector of random variables with

$$\mathbb{E}[\varepsilon|\mathbf{X}] = \mathbf{0}_n \quad \mathbb{V}\text{ar}[\varepsilon|\mathbf{X}] \equiv \mathbb{E}[\varepsilon\varepsilon^\top|\mathbf{X}] = \mathbf{W}^{-1}$$

allowing for

- non-constant variance
- correlated

residual errors.

Weighted least squares (cont.)

We estimate using *weighted least squares* (WLS)

$$\hat{\beta} = \arg \min_{\beta} (\mathbf{y} - \mathbf{X}\beta)^{\top} \mathbf{W} (\mathbf{y} - \mathbf{X}\beta)$$

that is, $\hat{\beta}$ solves

$$\mathbf{X}^{\top} \mathbf{W} (\mathbf{y} - \mathbf{X}\beta) = \mathbf{0}_p$$

so that

$$\hat{\beta} = (\mathbf{X}^{\top} \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^{\top} \mathbf{W} \mathbf{y}.$$

Unbiasedness:

$$\begin{aligned}\mathbb{E}[\hat{\beta}|\mathbf{X}] &= \mathbb{E}\left[(\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{Y} \middle| \mathbf{X}\right] \\ &= (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbb{E}[\mathbf{Y}|\mathbf{X}] \\ &= (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} (\mathbf{X} \beta) \\ &= (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{W} \mathbf{X}) \beta \\ &= \beta.\end{aligned}$$

Variance:

$$\begin{aligned}\mathbb{V}\text{ar}[\hat{\beta}|\mathbf{X}] &= \mathbb{V}\text{ar} \left[(\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{Y} \middle| \mathbf{X} \right] \\&= (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbb{V}\text{ar}[\mathbf{Y}|\mathbf{X}] \left\{ (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \right\}^\top \\&= (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{W}^{-1} \left\{ \mathbf{W} \mathbf{X} (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \right\} \\&= (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \left\{ \mathbf{W} \mathbf{X} (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \right\} \\&= (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{W} \mathbf{X}) (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \\&= (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1}.\end{aligned}$$

Weighted least squares (cont.)

Often, $\mathbf{W}$ is taken to be a *diagonal* matrix

$$\mathbf{W} = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix}$$

so w_i represents the reciprocal variance of ε_i .

This model allows for non-constant variance. In general $\mathbf{W}$ can be a function of the predictors, to allow for variances that depend on the predictors.

In the case of random predictors, we have for OLS that

$$\mathbb{V}\text{ar}[\hat{\beta}|\mathbf{X}] = \mathbb{E}[(\mathbf{X}^\top \mathbf{X})]^{-1}$$

and for WLS that

$$\mathbb{V}\text{ar}[\hat{\beta}|\mathbf{X}] = \mathbb{E}[(\mathbf{X}^\top \mathbf{W} \mathbf{X})]^{-1}$$

Suppose that for binary treatment Z which is randomly allocated with probability p , the model for outcome Y is

$$Y = \beta_0 + \beta_1 Z + \epsilon$$

This is a least squares problem with random predictor Z . We have that

$$\mathbf{x}^\top \mathbf{x} = \begin{bmatrix} 1 & z \\ z & z^2 \end{bmatrix} = \begin{bmatrix} 1 & z \\ z & z \end{bmatrix}$$

as $z^2 = z$.

To perform the estimation, we use OLS

$$\arg \min_{(\beta_0, \beta_1)} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 z_i)^2$$

and, differentiating, we obtain the two score equations

$$\beta_0 : -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 z_i) = 0$$

$$\beta_1 : -2 \sum_{i=1}^n z_i (y_i - \beta_0 - \beta_1 z_i) = 0.$$

We rewrite the two equations as

$$\sum_{i=1}^n y_i = n\beta_0 + \beta_1 \sum_{i=1}^n z_i$$

$$\sum_{i=1}^n z_i y_i = \beta_0 \sum_{i=1}^n z_i + \beta_1 \sum_{i=1}^n z_i^2.$$

Recalling that $z_i^2 = z_i$, and subtracting the second equation from the first, we have the equation,

$$\sum_{i=1}^n y_i - \sum_{i=1}^n z_i y_i = \left(n - \sum_{i=1}^n z_i \right) \hat{\beta}_0.$$

Hence, combining the sums, we have

$$\sum_{i=1}^n (1 - z_i) y_i = \sum_{i=1}^n (1 - z_i) \hat{\beta}_0.$$

and hence, provided the right-hand side sum is not zero,

$$\hat{\beta}_0 = \frac{\sum_{i=1}^n (1 - z_i) y_i}{\sum_{i=1}^n (1 - z_i)}.$$

which coincides with the moment-based estimator $\hat{\mathbb{E}}[Y(0)]$.

Plugging this into the second equation, we have

$$\sum_{i=1}^n z_i y_i = \hat{\beta}_0 \sum_{i=1}^n z_i + \beta_1 \sum_{i=1}^n z_i$$

and hence, provided the sum in the denominator is not zero,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n z_i y_i}{\sum_{i=1}^n z_i} - \hat{\beta}_0.$$

This coincides with the $\hat{\delta}$ estimator of the ATE.

Experimental study: estimation using OLS (cont.)

With Z random, we have that the asymptotic variance of the estimator is

$$\begin{aligned}\sigma^2 \mathbb{E}[\mathbf{X}^\top \mathbf{X}]^{-1} &= \sigma^2 \begin{bmatrix} 1 & \mathbb{E}[Z] \\ \mathbb{E}[Z] & \mathbb{E}[Z] \end{bmatrix} \\ &= \sigma^2 \begin{bmatrix} 1 & p \\ p & p \end{bmatrix}^{-1} \\ &= \sigma^2 \frac{1}{p(1-p)} \begin{bmatrix} p & -p \\ -p & 1 \end{bmatrix}\end{aligned}$$

Therefore, looking at the $(2, 2)$ element, we have that

$$\lim_{n \rightarrow \infty} n \text{Var}[\hat{\beta}_1] = \frac{\sigma^2}{p(1-p)}.$$